An Interpretable Approach to Fake News Detection

Caio Brighenti
Department of Computer Science, Colgate University

April 21, 2020

Abstract

The text of your abstract. 200 or fewer words.

Keywords:

1 Introduction

Propaganda has long been a tool of political influence, but in recent years it has taken a new online form: fake news. Fake news, once a buzzword on the internet, is now at the center of global politics, one of the most common bigram in the lexicon of United States president Donald Trump. After the term gained prominence in Trump's presidential campaign in 2016, it exploded into public conciousness, earning the distinction of Webster-Collins' "Word of the Year" in 2017.[1] As the current presidential race unfolds, fake news has returned to the center of the conversation, with major social media companies facing scrutiny of their misinformation policy. This phenomenon is also not a distinctly American problem–investigate reporting both during and after the 2018 Brazilian president election demonstrated that more than XX% of news articles shared on the popular messaging service WhatsApp were fake news.[2]

Fake news is not only politically significant but also dangerously tempting. Studies have shown false content propagates faster through social media than real content.[3] Blatantly false or exaggerated rhetoric can even lead to violent action, as demonstrated by the "Pizzagate" incident in which a man stormed a D.C. pizzeria with an AR-15, having been convinced by false and unverified information that a pedophile ring operated out of the restaurant's basement.[4] Fake news can also be incredibly easy to create. In 2019, a group of researchers at the Allen Institute for Artificial Intelligence published text generation model able to produce fake news.[5] In a troubling conclusion, the researchers found that state-of-the-art fake news detection systems struggled more with identifying fake news produced by their systems than actual fake news.[6] Fake news is thus easy to create, spreads quickly, and is hard to detect, a dangerous combination making it a serious threat to civic society.

Given the danger that fake news poses, machine learning and natural language processing researchers have devoted significant attention to the problem of fake news classi-

1

2

3

4

5

6

fication. However, previous attempts at fake news classification overwhelmingly rely on highly complex models suffering from the "black box" problem. As a result, these models lack interpretability and do not allow us to reach new conclusions about the nature of fake news. In order to begin closing this gap, this research adopts an interpretable approach, with the overall objective of a producing a fake news classification model with comparable accuracy to state of the art models without compromising interpretability.

PARAGRAPH SUMMARIZING FINDINGS

2 Prior Work

Since the 2016 U.S. presidential election, fake news has been a frequent topic of natural language processing research. There are countless examples of papers approaching fake news classification or closely related problems, but from slightly different angles. This section summarizes the prior work in fake news detection, clarifying the different categories of methods. In general, previous works in fake news detection differ in three major ways: 1) the scale of the predicted variable, the information used as features, and the type of model. With respect to scale, any fake news detection model falls into one of three levels of granularity: 1) claim level, 2) source level, and 3) article level. These levels of analysis describe the response variable being predicted.

Claim level approaches attempt to determine whether a given claim, usually one or several sentences, is true or intentionally misleading. [7] Given that claim-level approaches must make a judgement based on only a short amount of text, researchers often adopt a fact-checking strategy This strategy, also known as "truth discovery," assumes that a sentence's claims can be gramatically isolated and checked against a database of established claims.[8] A natural application for claim-level models is social media, most commonly Twitter, where little is known about the author and only a very limited amount of text is available. However, claim-level approaches have serious limitations, often struggling with the complex sentences journalists or other writers typically employ.[9] Aditionally, they rely

[7] Examples of claim-level approaches include ...

[8] strube p. 2

[9] strube 2

entirely on a complete knowledge base, which must be constantly expanded and updated, clearly a difficult task.

Source level approaches attempt to classify whether a speaker or entire news source consitently publish misinformation. The intuition behind these approaches is that speakers or sources that have published misinformation in the past are likely to continue to do so. An example of a source-level approach is the popular browser extension "BS Detector,"[10] which classifies articles on a fine-grained scale of veracity by checking the source's status in a database of news sources and their reliability. A source's history of misinformation can also be used as a predictor in claim-level or article level approaches. Kirilin and Strube, for instance, create *Speaker2Credit*, a metric of speaker credibility, and show how it can improve the performance of fake news detection models when used as an input.[11]

Article level approaches have received the most attention in the work on fake news detection. This is logical, given that fake news tends to take the form of articles, peddling misinformation in the article text while posing as a legitimate source. Article-level approaches also benefit from a rich list of predictors to choose from, including not only the article's content but all relevant metadata. Kai Shu et. al, for instance, build an article-level model using linguistic and visual components of the article content, the social context around it—including information on the user that posted it, the post itself, responses to it, and the social network of the poster—, as well as spatiotemporal information capturing when and where the article and responses were to it were posted from.[12]

The work of Shu et. al is an example of the overwhelming number predictors available to researchers working in fake news detection, resulting in a diversity of approaches with respect to feature selection. Melanie Tosik et. al, for instance, employ only hand-crafted features capturing the similarity between an article's title and text in a two-stage ensemble classifier modeling whether an article's body agrees with its headline.[13] Sonam Tripathi and Tripti Sharma demonstrate the effectiveness of parts of speech tagging—also known as grammatical tagging—in document classification problems, the general category of nat-

[10]

[11] Kirilin and Strube

[12] Shu et al

[13] Tosik et al

ural language processing that article-level fake news detection falls under.[14] Ramy Baly et al. employ a breadth of features to model factuality and bias of news sources, using features covering the content of articles, the source's Wikipedia and Twitter pages, the structure of the URL, and the source's web traffic. [15]

Independent of level of analysis or choice of predictors, researchers overwhelmingly choose to use complex deep neural networks. Ajao et. al, for instance, use a "hybrid of convolutional neural networks and long-short term recurrent neural network models" to classify Tweets as true or false based on their text content.[16] The dominance of deep learning approaches is visible in an extensive survey on fake news detection done by Ray Oshikawa and Jing Qian. [17] The pair's section on machine learning models dedicates a total of three sentences to "Non-Neural Network Models," compared to seven paragraphs. focusing on neural networks.

While deep learning approaches can produce highly accurate models that consistently succeed in identifying misinformation, they also suffer from a lack of interpretability. This is often referred to as the "black-box" problem, meaning that the inputs and outputs of these models are perfectly clear, but the steps that the model takes to reach the output are completely invisible. This has been identified as a limitation of the the work on fake news detection thus far.[18] Oshikawa and Qian, at the conclusion of their extensive survey, declare that "we need more logical explanation for fake news characteristics," highlighting the need for models that can teach us something about fake news.

Approaches that focus on interpretability are rare, but do exist. From the deep learning perspective, Nicole O'Brien et al. employ post-hoc variable importance to their text-based deep learning model, identifying the words that are most predictive of fake and real news.[19] Their approach, however, does not interpret the results, but instead merely demonstrates the feasiblity of the technique. Furthermore, this method reveals only information about *specific* words, as opposed to *types* of words. More applicable is the work of researchers

[14] tripathi

[15]

[16] ajao et al

[17] Oshikawa

[18] shu, O'Brien

[19] O'Brien et al

who both use features that describe the semantic properties of the text in general, use interpretable models, and extensively document their results. The best examples of this type of work are the works of Benajmin Horne et al. and Mauricio Gruppi et al., both of which employ features capturing the complexity, style, and psychology of fake news, and display precisely how fake and real news differs in each of the variables used.[20]

3 Methodology

This paper seeks to contribute to the small literature of interpretable fake news detection, by following the methodology of Horne et al. and Gruppi et al., leveraging features that describe textual properties of fake news, applying non-neural network models and focusing heavily on interpretation of results. This section details the precise methodology employed, discussing the dataset used, the feature engineering process, outlier removal, and models applied.

There are many datasets freely available for fake news detection, but many suffer quality limitations. Oshikawa and Qiang outline 12 requirements for a quality fake news dataset, expanding a 9-point list originally by Rubin et al.: 1. Availability of both truthful and deceptive instances; 2. Digital textual format acessibility; 3. Verifiability of ground truth; 4. Homogeneity in lengths; 5. Homogeneity of writing mattes; 6. Predefined timeframe; 7. The manner of news delivery; 8. Pragmatic concerns; 9. Language and culture; 10. Easy to create from raw data; 11. Fine-grained truthfulness; 12. Various sources or publishers.[21] While no dataset currently exists that meets all 12 criteria, the article-level *FakeNewsNet* (FNN) dataset meets most.[22]

FNN is an article-level dataset that includes the title and body of each article (2), each of which have been profesionally fact-checked and labeled as true or false by Politifact or Gossicop(1, 3). FNN provides both political and celebrity news articles, but this paper chooses to use only the political articles in order to maintain a roughly consistent corpus (4,5,7). These articles are all in English, largely center around American politics, and

[20]Horne, Gruppi

[21]Oshikawa, Rubin

[22]FNN

come from a variety of sources (9, 12). Finally, there is little work needed to obtain the dataset, as FNN provides an API to quickly obtain the body and title of each article (10). The biggest limitation for FNN is that it lacks a fine-grained scale of truth, labeling only as binary true/false (11). However, given that it meets the majority of the criteria, and contains XXX observations, it is overall a good fit for this paper.

4 Results

5 Discussion

6 Future Work